Epilepsia Open® The Open Access Journal of the International League Against Epilepsy

Open Access

FULL-LENGTH ORIGINAL RESEARCH

Simulating clinical trials with and without intracranial EEG data

*Daniel M. Goldenholz, *†Joseph J. Tharayil, ‡Rubin Kuzniecky, §Philippa Karoly, *William H. Theodore, and §Mark J. Cook

Epilepsia Open, 2(2):156–161, 2017
doi: 10.1002/epi4.12038

Dr. Daniel M. Goldenholz focuses his research on understanding the natural variability in epilepsy as well as developing better tools to prevent SUDEP.

Summary

Objective: **It is currently unknown whether knowledge of clinically silent (electrographic) seizures improves the statistical efficiency of clinical trials.**

Methods: **Using data obtained from 10 patients with chronically implanted subdural electrodes over an average of 1 year, a Monte Carlo bootstrapping simulation study was performed to estimate the statistical power of running a clinical trial based on (1) patient-reported seizures with intracranial electroencephalogram (icEEG) confirmation, (2) all patient-reported events, or (3) all icEEG-confirmed seizures. A "drug" was modeled as having 10%, 20%, 30%, 40%, and 50% efficacy in 1,000 simulated trials each. Outcomes were represented as percentage of trials that achieved $p < 0.05$ using Fisher's exact test for 50% responder rates (RR50) and the Wilcoxon rank-sum test for median percentage change (MPC).**

Results: **At each simulated drug strength, the MPC method showed higher power than RR50. As drug strength increased, statistical power increased. For all cases except RR50 with drug of 10% efficacy, using patient-reported events (with or without icEEG confirmation) was not as statistically powerful as using all available intracranially confirmed seizures ($p < 0.001$).**

Significance: **With simulation, this study demonstrates that additional accuracy in seizure detection using chronically implanted icEEG improves statistical power of clinical trials. Newer invasive and noninvasive seizure detection devices may have the potential to provide greater statistical efficiency, accelerate drug discovery, and lower trial costs.**

KEY WORDS: **Clinical trial, Monte Carlo, Intracranial EEG, Biostatistics.**

Background

Epilepsy clinical trials suffer from numerous error sources. Usually, they are performed on outpatients over several months.[1] The primary outcome measure is derived from patient-reported seizure counts, normalized by each individual's baseline, and compared between a placebo and a treatment arm. Unfortunately, patient-reported clinical seizures may substantially underestimate seizures recordable intracranially, perhaps by a factor of 10 to 1.[2] In addition, patient diaries overreport other events as seizures.[3]

Modern clinical trials have been affected adversely by skyrocketing costs (Pharma 2015) and a steadily rising "placebo effect."[4] For instance, one recent trial reported placebo effects as high as 40%.[5] Higher "placebo effects" decrease trial efficiency, increasing costs.[6]

Accepted December 21, 2016.

*Clinical Epilepsy Section, NINDS, NIH, Bethesda, Maryland, U.S.A.; †Duke University, Department of Biomedical Engineering, Durham, North Carolina, U.S.A.; ‡NYU Epilepsy Center, New York University, New York, New York, U.S.A.; and §University of Melbourne, Fitzroy, Victoria, Australia

Address correspondence to Daniel Goldenholz, National Institutes of Health, NINDS, Clinical Epilepsy Section, CNP, DIR, 10 Center Drive, 10-CRC, Room 5S-207, MSC 1408, Bethesda, MD 20892-0001, U.S.A. E-mail: daniel.goldenholz@nih.gov

Key Points

- More seizures are recordable with intracranial EEG than merely the clinically reported ones
- This study simulated the possibility of running clinical trials with and without the intracranial EEG–recorded seizures
- The main finding was that using the intracranial EEG–recorded seizures significantly increased the statistical power of the simulated trials

It may be prudent to consider methodologies for improving trial efficiency using modern technology. Although we expect a larger number of events to improve the efficiency of clinical trials, such improvement has never been rigorously investigated for epilepsy subjects. Furthermore, quantifying the magnitude of the expected efficiency gains will guide future trial strategies, for instance, the use of invasive or semi-invasive, subscalp recording electrodes for human trials. It is both feasible and safe to implant subdural electrodes chronically over many months.[2] Implantation vastly increases detection sensitivity and specificity, thereby providing an accurate seizure catalog. Although a very small number of patients have actually had chronic subdural electrodes placed over months to years, there is a rich data set from which to derive simulations about hypothetical situations. One such hypothetical is this: does knowledge of both intracranially recorded and clinically reported seizures increase the efficiency of a clinical trial?

Methods

We simulated a randomized clinical trial, based on a recent trial of 15 patients with chronically implanted subdural electrodes, using custom software in Matlab (R2015b) and R (3.2.3). Data included intracranial electroencephalogram (icEEG), patient-reported diaries, as well as audio recordings at the time of electrographic seizures.[2] Using methods previously described,[2,7] events were annotated into several subtypes: (1a) clinically manifested with correlated EEG ictal activity, (1b) clinically unreported seizures with audio and EEG features of a clinical seizure, (2) those with EEG ictal pattern matching those of subtype 1a/1b events but lacking confirmation of clinical manifestation, (3) EEG with seizure-like characteristics but differing from subtype 1a and 1b events and without confirmation of clinical manifestation, and (4) clinically reported seizures in the complete absence of electrographic confirmation.

Because of small sample size, we employed a form of within-block bootstrapping[8] to produce virtual patient data for a full trial period. First, a random patient (uniform distribution, with replacement) from the available NeuroVista patients was selected. This means that each patient had an equal likelihood of being chosen and could be chosen multiple times. Then, a random start day (uniformly distributed, with replacement) was selected from all possible start days in that patient's diary. All days (other than the final 6) were possible start days. One week (7 days) of contiguous seizure data was obtained starting at that day from the patient diary. The process of collecting 1 week of seizure data was repeated 20 times. Thus, each virtual patient was developed independently from a single patient, using a set of randomly chosen weeks. Because the choice of start times was very large, and 20 such choices were needed, the total number of unique possible combinations was 4.13×10^{51}. A set of 200 virtual patients was generated for each virtual trial (Fig. 1). Each virtual trial was 5 months long: 2-month baseline and 3-month test period.

Each virtual patient contributed to three seizure diaries: (A) patient-reported seizures that were electrographically confirmed (i.e., subtype 1a), (B) patient-reported events (i.e., subtypes 1a and 4), and (C) all electrographically confirmed seizures (subtypes 1a, 1b, and 2), captured by the

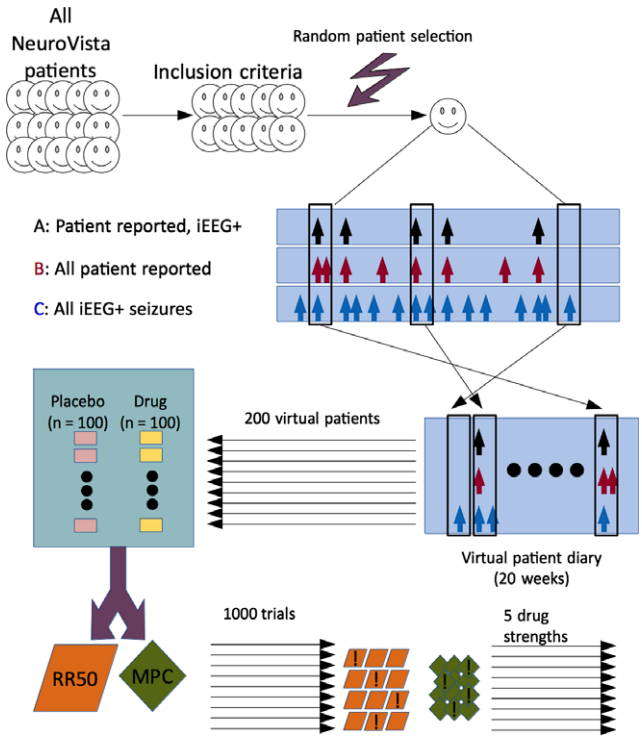

Figure 1.
Flow diagram. A subset of NeuroVista patients was used based on inclusion criteria. To generate one virtual patient, first a random patient was selected. Then, 20 windows of duration 1 week were selected, with random start times. The process was repeated 200 times for each clinical trial, which comprised 100 placebo patients and 100 drug patients. The trial was analyzed with the 50% responder rate (RR50) and the median percent change (MPC) methods. Some trials were successful (denoted with an exclamation mark) at distinguishing drug from placebo. For each of five drug strengths, 1,000 trials were simulated.
Epilepsia Open © ILAE

intracranial electrodes only. From the perspective of patient reporting, category A represented all "true positives," B represented "true positives" and "false positives," and C represented "true positives" and "false negatives." All patients were assumed to complete the trial (i.e., no dropout), simulating the optimal situation. Each of the original 15 patients was included only if the expected duration needed to obtain at least one patient-reported event (category A) was less than the size of the baseline (2 months). For each week selected using the bootstrap, all three categories derived from the original data were retained in the virtual patient, thus preserving the temporal relationship between categories within 1-week intervals.

Based on recent data,[6] placebos were modeled as natural variability alone; therefore, no simulated adjustments were required for the existing seizure counts. Drugs were simulated using several assumptions: (1) patients in the trial had 100% compliance; (2) the drugs were equally effective in all virtual patients; (3) the efficacy of the drugs was stable throughout the exposure for each virtual patient; (4) the efficacy of the drugs was memoryless, that is, the effect of the drug at any time was independent of any other time that the drug was used; and (5) the drug had a fixed ability to decrease the total percentage of seizures potentially experienced. On the basis of these assumptions, the drug would require a single "efficacy" parameter E, which would represent the percentage of time that the drug would prevent any given seizure that was about to occur. With this model, a drug was modeled by removing the ith seizure with probability E during the testing period.

Two very commonly used trial outcome measures were considered: 50% responder rates (RR50) compared with Fisher's exact test, and median percentage changes (MPC) compared with the Wilcoxon rank-sum test. In both cases, these tests are used to accept or reject the null hypothesis that the drug and placebo arms are equivalent. If rejected at the $p < 0.05$ level, a trial is typically considered "successful."

Drug efficacies 10%, 20%, 30%, 40%, and 50% were tested 1,000 times each with the bootstrapped trials in each of the three categories (A, B, and C) for a total of 15,000 simulated trials. Because each virtual patient contributed three categories, these trials required $5{,}000 \times 200 = 1{,}000{,}000$ virtual patients. Each trial had six outcome p values because of the two outcome measures and the three diary categories considered. A trial "success" was defined as $p < 0.05$; thus, each trial contributed six binary success variables.

A Wilcoxon rank-sum test compared the sets of 1,000 p values obtained in two contrasts: category A versus category C, and B versus C. This comparison was performed for each of the five drug strengths and both outcome methods (RR50 and MPC), resulting in $2 \times 5 \times 2$ comparisons. The 20 tests were adjusted for multiple comparisons using the Bonferroni correction.

Results

On the basis of inclusion criteria, 10 patients from the NeuroVista trial were included for simulation. The seizure diaries ranged from 7 to 24 months in duration (median 12).

Table 1 shows the mean and standard deviations of the 1,000 p values obtained in each of the simulated situations. As expected, larger drug strengths dramatically decrease the p values in all cases. Also as expected, p values obtained from MPC were consistently lower than RR50.

Figure 2 shows the percentage of the 1,000 trials that met the $p < 0.05$ level of significance, which is equivalent to an estimate of the statistical power. Again, the increasing drug strength shows increasing statistical power for all cases. Also again, the MPC method consistently demonstrates superior statistical power.

The category C events always showed lower p values than category B events ($p < 0.001$, Wilcoxon rank-sum test, corrected), and category C also always showed lower p values than category A ($p < 0.001$, Wilcoxon rank-sum test, corrected). The two exceptions were that the RR50 method for strength 10% did not show a statistical difference between B and C ($p > 0.05$, Wilcoxon rank-sum test, corrected) and that it did show a significant difference between A and C ($p > 0.05$, Wilcoxon rank-sum test, corrected).

Discussion

Our study found that using all electrographically confirmed seizures affords statistically superior power compared to using only clinically reported events. The finding was true across a series of drug strengths, two different standard methods of calculating the outcome of a clinical trial, and with or without clinically reported false positives.

Implicit in our study is the assumption that drugs capable of reducing subclinical seizures are beneficial to patients. The relevance of subclinical seizures for assessing the efficacy of antiepileptic drugs is inconclusive, yet there is likely to be a patient-specific relationship between clinical and subclinical seizures. Subclinical seizures have been shown to be reliable indicators of the epileptogenic zones,[9–11] suggesting a common pathology with clinical seizures. It is possible that subclinical discharges have a similar generating mechanism to clinical seizures.[12] Moreover, subclinical discharges appear to have cognitive impacts as well, even when controlling for lesion, drug, and duration of epilepsy.[13] Given the available evidence, we believed it reasonable to allow simulated drugs to have equal efficacy across clinical and subclinical seizures. Therefore, measuring the effect of drugs on both types of events may be clinically relevant.

One might think that the results of this study are expected. It would seem "obvious" that having more events would certainly result in greater statistical power. However, several

Table 1. Expected p values. For each drug strength (in %), the expected p values mean ± standard deviation are shown

Drug	A—RR50	A—MPC	B—RR50	B—MPC	C—RR50	C—MPC
10	0.503 ± 0.322	0.367 ± 0.308	0.507 ± 0.323	0.352 ± 0.303	0.509 ± 0.346	0.231 ± 0.268
20	0.293 ± 0.297	0.133 ± 0.205	0.295 ± 0.298	0.115 ± 0.194	0.189 ± 0.256	0.025 ± 0.082
30	0.088 ± 0.161	0.019 ± 0.061	0.081 ± 0.156	0.014 ± 0.054	0.022 ± 0.079	0.000 ± 0.004
40	0.010 ± 0.035	0.001 ± 0.005	0.008 ± 0.031	0.001 ± 0.002	0.000 ± 0.001	0.000 ± 0.000
50	0.000 ± 0.005	0.000 ± 0.000	0.000 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

A, clinically reported seizures with electrographic confirmation; B, all clinically reported events; C, all electrographically confirmed seizures; RR50, 50% responder rate, tested with Fisher's exact test; MPC, median percentage change, tested with the Wilcoxon rank-sum test.

things could have happened differently. First, if there was a wide degree of variability in the subclinical events coupled with a low variability in the clinical events, then including subclinical events would result in *lower* statistical power (Appendix S1). Second, if sufficient subtype 4 (false-positive) events were present, then the statistical power of category B would be degraded. In reality, the precise long-term relationship between intracranially confirmed seizures and patient-reported events remains largely unknown. The data available on this relationship come from only one study—the NeuroVista trial itself.[2] As a consequence, although in hindsight the results appear intuitively satisfying, the outcome was not a forgone conclusion.

The limitations of this study are based on the assumptions made. There is a very small sample of original patients available who have ever been studied longitudinally with intracranial electrodes over months to years. Because thousands of seizures were recorded over this time, it was feasible to generate virtual patients using a form of bootstrapping. However, the results may not generalize to all patients and all forms of epilepsy because of the small number of patients studied here. Indeed, it is almost certainly true that a larger sample of patients with longitudinal intracranial recordings would enrich this simulation with a greater heterogeneity of seizure frequencies, variability, clustering, and so on. Moreover, by necessity, the assumption of true independence between virtual patients will sometimes be violated to a limited extent, because some virtual patients will have been generated from the same "true" patient, though from differing portions of the seizure diary. Consequently, our conclusions must be viewed as merely a first approximation based on our limited data available.

Additional assumptions were made about the placebo effect, namely, that the placebo-exposed patients experienced no change in their typical seizures, and any "response" was primarily an artifact of natural variability.[6] Although the responses obtained from natural variability are typical for clinical trials,[14] this assumption has not been fully proven. Although other models could have been added to our simulation, such as regression-to-the-mean and psychological effects, these influences have never been formally quantified in epilepsy. Therefore, we elected to use a model of placebo that has been quantified (i.e., natural variability) and avoid the additional unproven assumptions required for additional factors to be included as well. Indeed, it is possible that at least a small portion of variability seen in the NeuroVista data may reflect medication changes, though medications were for the most part stable throughout that trial. Similarly, assumptions about drug efficacy are speculative at best, though they may represent at least a reasonable first approximation.[15] Our model assumed 0% dropout, which is obviously unrealistic, though lower dropout rates might be expected in highly invested patients with implanted devices. However, dropout will simply decrease the statistical power of a study, so these results can be used as a guide for a best-case scenario. Finally, the assumption that electrographically captured events represent all seizures may be naïve—it is unknown how much intracranial electrode coverage would identify all electrographic seizures with 100% sensitivity. In the case of NeuroVista, a set of 16 electrodes was used, covering the area expected to be most likely the epileptogenic zone.

The implication of this study is that if a set of patients with chronically implanted intracranial electrodes was available, clinical trials using data obtained from these patients would have higher statistical power. Such patients may become available incidentally as devices with intracranial recording capabilities become more common, such as the RNS and DBS systems. Extracranial "subscalp" systems are also being developed, which will provide an alternative means of collecting similar data. Because of the higher statistical power, the number of patients required for a clinical trial could be decreased—in some cases by a dramatic amount. The improved power therefore translates to lower costs, which could accelerate drug discovery.[16] Higher statistical power could also decrease exposure to subtherapeutic doses of medications, lowering the risk of Sudden Unexpected Death in Epilepsy (SUDEP).[17] It is likely that with improving detection and prediction algorithms more patients will be willing to have chronic subdural electrodes implanted in the coming years.[2,18] Similarly, if less invasive techniques, such as wearable biosensors,[19–22] were to achieve high reliability and accuracy, they too would be predicted to obtain higher statistical power than self-reported seizure diaries. Indeed, any technology

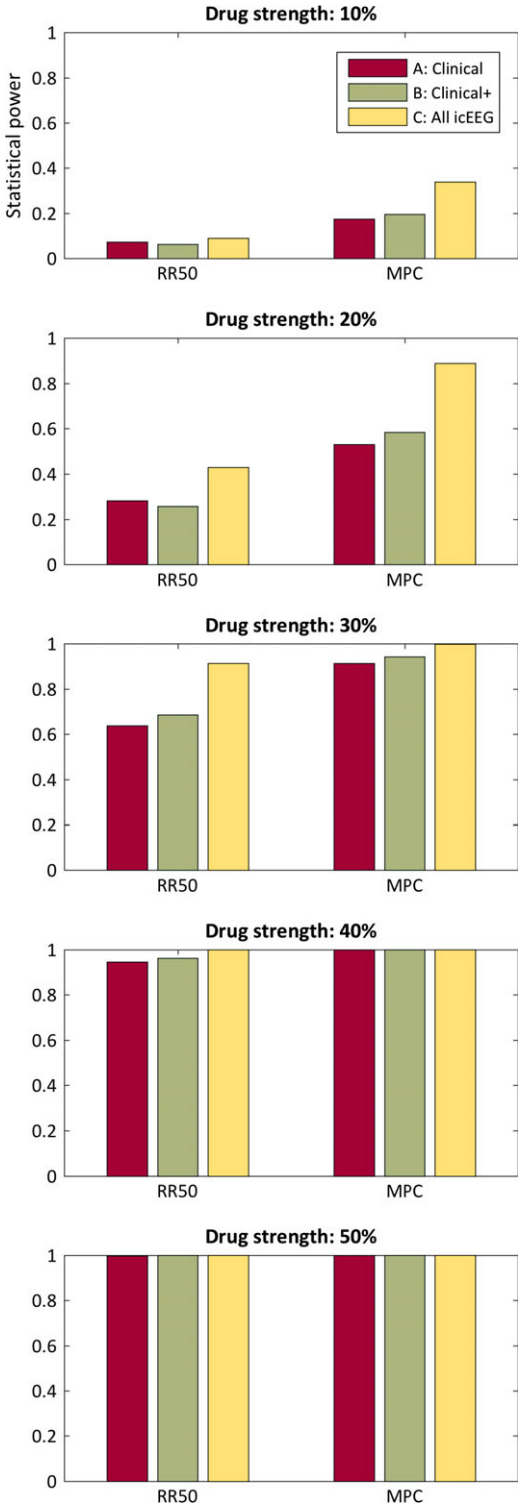

Figure 2.
Statistical power. Each bar represents the percentage of 1,000 trials that achieved $p < 0.05$ statistical significance, which is an estimate of statistical power. **A**, clinically reported seizures with electrographic confirmation; **B**, all clinically reported events; **C**, all electrographically confirmed seizures; RR50, 50% responder rate, tested with Fisher's exact test; MPC, median percentage change, tested with the Wilcoxon rank-sum test.

that increases the number of true detections of seizures could improve the landscape of clinical epilepsy trials by lowering costs, shortening trials, and perhaps even saving lives.

Acknowledgments

This study was funded by the National Institute of Neurological Disorders and Stroke Intramural Research Program. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Disclosure

The authors declare not conflicts of interest. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

References

1. Perucca E. What clinical trial designs have been used to test antiepileptic drugs and do we need to change them? *Epileptic Disord* 2012;14:124–131.
2. Cook MJ, O'Brien TJ, Berkovic SF, et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *Lancet Neurol* 2013;12:563–571.
3. Fisher RS, Blum DE, DiVentura B, et al. Seizure diaries for clinical research and practice: limitations and future prospects. *Epilepsy Behav* 2012;24:304–310.
4. Rheims S, Perucca E, Cucherat M, et al. Factors determining response to antiepileptic drugs in randomized controlled trials. A systematic review and meta-analysis. *Epilepsia* 2011;52:219–233.
5. French JA, Krauss GL, Wechsler RT, et al. Perampanel for tonic-clonic seizures in idiopathic generalized epilepsy A randomized trial. *Neurology* 2015;85:950–957.
6. Goldenholz DM, Moss R, Scott J, et al. Confusing placebo effect with natural history in epilepsy: a big data approach. *Ann Neurol* 2015;78:329–336.
7. Cook MJ, Karoly PJ, Freestone DR, et al. Human focal seizures are characterized by populations of fixed duration and interval. *Epilepsia* 2015;57:359–368.
8. Ju H. Moving block bootstrap for analyzing longitudinal data. *Commun Stat Theory Methods* 2015;44:1130–1142.
9. Velkey A, Siegler Z, Janszky J, et al. Clinical value of subclinical seizures in children with focal epilepsy. *Epilepsy Res* 2011;95:82–85.
10. Sperling MR, O'Connor MJ. Auras and subclinical seizures: characteristics and prognostic significance. *Ann Neurol* 1990;28:320–328.
11. Farooque P, Duckrow R. Subclinical seizures during intracranial EEG recording: are they clinically significant? *Epilepsy Res* 2014;108:1790–1796.
12. Davis KA, Ung H, Wulsin D, et al. Mining continuous intracranial EEG in focal canine epilepsy: relating interictal bursts to seizure onsets. *Epilepsia* 2016;57:89–98.
13. Glennon JM, Weiss-Croft L, Harrison S, et al. Interictal epileptiform discharges have an independent association with cognitive impairment in children with lesional epilepsy. *Epilepsia* 2016;57:1436–1442.
14. Goldenholz DM, Goldenholz SR. Response to placebo in clinical epilepsy trials—old ideas and new insights. *Epilepsy Res* 2016;122:15–25.
15. French JA, Kugler AR, Robbins JL, et al. Dose-response trial of pregabalin adjunctive therapy in patients with partial seizures. *Neurology* 2003;60:1631–1637.
16. PhRMA. Profile Biopharmaceutical Research Industry. 2015. http://www.phrma.org/sites/default/files/pdf/2015_phrma_profile.pdf.
17. Ryvlin P, Nashef L, Lhatoo SD, et al. Incidence and mechanisms of cardiorespiratory arrests in epilepsy monitoring units (MORTEMUS): a retrospective study. *Lancet Neurol* 2013;12:966–977.

18. Morrell MJ. Responsive cortical stimulation for the treatment of medically intractable partial epilepsy. *Neurology* 2011;77:1295–1304.
19. Poh M-Z, Loddenkemper T, Reinsberger C, et al. Convulsive seizure detection using a wrist-worn electrodermal activity and accelerometry biosensor. *Epilepsia* 2012;53:e93–e97.
20. Kramer U, Kipervasser S, Shlitner A, et al. A novel portable seizure detection alarm system: preliminary results. *J Clin Neurophysiol* 2011;28:36–38.
21. Velez M, Fisher RS, Bartlett V, et al. Tracking generalized tonic-clonic seizures with a wrist accelerometer linked to an online database. *Seizure* 2016;39:13–18.
22. Szabó CÁ, Morgan LC, Karkar KM, et al. Electromyography-based seizure detector: preliminary results comparing a generalized tonic-clonic seizure detection algorithm to video-EEG recordings. *Epilepsia* 2015;56:1432–1437.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Why the results are not obvious.